基于社区增量自适应爬虫研究

马 睿

(暨南大学 信息科学技术学院,广东 广州 510632)

摘 要：在分析传统的网络蜘蛛搜索特点的基础上,充分利用 Web 资源分布的特点,提出了基于在线增量自适应算法的搜索策略。该算法一方面避免了过早陷入 Web 搜索最优子空间的陷阱;另一方面不断对爬虫数据库更新,以提高其对链接主题的判断能力。通过对四所著名大学计算机网站做的搜索实验,表明新的算法可以有效地提高网络蜘蛛的搜索性能。

关键词：网络蜘蛛;搜索策略;在线增量自适应

中图分类号：TP391.3 **文献标识码**：A **文章编号**：1674-7720(2010)21-0045-04

Spider study based on incremental adaptive of community

MA Rui

(School of Computer Science and Technology, Jinan University, Guangzhou 510632, China)

Abstract: After abstracting the characteristics of traditional Web spider and making full use of characteristics of Web resources distribution, the paper proposes community incremental adaptive algorithm search strategy. On the one hand, the algorithm avoids premature falling into the trap of Web optimal subspace. On the other hand, it constantly updates the database to enhance its ability to link the subject to judge. The search experiment of four famous university's computer Website shows that the new algorithm can improve the search properties of Web spider.

Key words: Web spider; search strategy; on-line incremental adaptive

随着 Internet 的快速发展,Web 上的信息资源也呈指数级增长,搜索引擎已经成为网络用户获取各种信息的必备工具。对于搜索引擎来说,要抓取互联网上的所有信息几乎是不可能的,从公布的数据来看,容量最大的搜索引擎也只不过是抓取了整个网络信息的 40%左右。传统的搜索引擎(如 Google、Baidu、Yahoo 等)大多数都是面向所有信息的搜索引擎,是一种通用搜索引擎。这种通用搜索引擎已经不能满足特定用户更深入的查询需求,他们对信息的需求往往是面对特定领域和特定主题的。面对挑战,适应特定人群需要的专业搜索引擎逐渐引起研究学者的重视。

主题网络蜘蛛是最近几年兴起的研究热点,它针对某个专门的领域进行搜索,以满足特定人群的个性化需求。网络蜘蛛研究的核心是解决页面和 URL 的主题相关性判别的问题,因此如何评价链接价值就成了网络蜘蛛爬进效率的关键。链接价值可以分为两类,即基于立即回报价值和基于未来回报价值。

立即回报价值算法是依据搜索时在线获得的文本或 Web 结构来对链接的页面重要程度进行预测,进而决定链接访问顺序。这类方法理论基础好,计算简单,在距离相关页面比较近的时候表现出良好的性能。但它很难反映 Web 的整体情况,网络蜘蛛在距离相关页面比较远的时候容易迷失方向。基于未来价值的算法利用 Web 上的信息分布在某种程度的相似性,对网络蜘蛛先进行训练,使其具有一些经验信息,对未来搜索具有一定的预测性。但其预测能力有限,而且需要用户选择种子集,搜索时不灵活,容易引起主题漂移[1-4]。

本文基于两类评价方法,提出了一种在线学习的自适应综合价值的网络蜘蛛搜索算法,利用 Web 资源分布的某些相似性和链接价值的关系,将立即价值和未来价值的评价方法相结合,在爬行过程中不断自身提高链接主题相关性的判断能力,从而改进网络蜘蛛的性能。

1 主题爬行策略

根据评价链接价值所采用的不同方法,现有的网络

蜘蛛的搜索策略分为两大类：基于立即回报价值评价的搜索策略和基于未来回报价值的搜索策略。本文采用基于内容评价的策略（基于立即价值）和基于巩固学习的搜索策略（基于未来价值）。

基于内容评价的搜索策略，主要是根据主题与链接文本“语义”的相似度来评价链接价值的高低。链接文本是指周围的说明文字和和链接 URLs 上的文字信息。相似度的评价一般采用下面的公式：

$$sim(q,p)=\frac{\sum_{k\in q\cap p}w_{kq}w_{kp}}{\sqrt{\sum_{k\in p}w_{kp}^2\sum_{k\in q}w_{kq}^2}} \quad (1)$$

其中，q 代表主题关键词集合，p 代表页面链接文本集合，W_{kp} 代表 d 中单词对某一主题的重要程度，W_{kp} 通常采用 tf×idf 公式计算。

基于巩固学习的搜索策略，巩固学习的优势在于能预测远期的回报价值（也称未来价值）。未来价值用 Q 来表示，这种方法的核心就是如何计算链接的 Q 价值。为此，搜索过程被分为训练和搜索两个阶段。训练阶段用巩固学习算法计算每个链接的 Q 价值，按价值的大小分为若干类，并用每一类中的文本信息训练一个 Naive Bayes 分类器；在搜索阶段，面对价值未知的链接，则根据链接文本，用 Naive Bayes 分类器计算链接落在每一类中的概率，并以这个概率为权值来计算链接的 Q 价值。因为 Q 价值反映的是未来的回报预测值，所以当搜索的页面与主题不相关时，网路蜘蛛也可以根据未来回报的预测值来确定正确的搜索方向。

该模型的核心就是如何计算链接的 Q 价值。Q 价值的计算公式[5]：

$$Q^{\lambda}(st,at)=(1-\lambda)(Q^{(1)}{}_t+\lambda Q^{(2)}{}_t+\cdots+\lambda^2 Q^{(3)}{}_t+\cdots) \quad (2)$$

其中 $Q^{(n)}{}_t$ 是各链接前瞻 n 步的折算积累回报，该 $Q(\lambda)$ 函数使用常量 $\lambda(0\leqslant\lambda\leqslant1)$ 来合并从不同前瞻距离中获得的回报。

2 在线增量自适应算法网络蜘蛛搜索策略

2.1 Web 资源分布和链接价值的关系

虽然整个网络资源的分布是无序的，但近年来的研究表明，与某一主题相关的页面以不同群聚群体的方式分散在网络中，把这些群体称为 Web 社区[6]。图 1 中显示了这种 Web 社区的分布关系。在网页的设计过程中不可能把所有相关的网页链接在一起，网页中只包含了极少部分与该主题相关的网页链接，这些资源信息一起构成了一个与某一主题相关的 Web 社区。在某一站点附近有很多紧密联系的站点，它们都能基本地反映某个主题。但在网页的发布过程中，可能会出现与之有一定关联但又与主题不相关的无关网页，这些无关网页在网络蜘蛛的爬行过程中会导致中心主题发生漂移。正是由于这些无关网页的“干扰”，使得网络蜘蛛在爬行的过程中随着时间的推移，爬行出来的链接会与最初链接的主题相关性差别越来越大，系统爬行到的网页也越来越少。这就要求在网络蜘蛛的爬行过程中，一方面能尽可能地覆盖所有相关网页；另一方面又要在爬行的过程中不断“更新”，以提高主题相关性的判断能力。这就要求网络蜘蛛在不同的阶段采用不同的搜索策略，同时不断“自我更新”，以提高爬行的效率和精度。

图 1 Web 信息资源分布特点示意图

2.2 算法思想

根据 Web 资源的分布信息，本文把网络蜘蛛爬行的过程人为地分为两个阶段：挖掘和探索。在 Web 社区内，由于和主题相关的网页比较多，立即价值比较大，这个时候就要求能尽快地挖掘 Web 社区内与主题相关的网页信息。这个时候适合选取注重发掘立即价值的搜索策略。而在 Web 社区之间，由于存在大量与主题无关的网页，这个时候要注重探索，尽可能地探索到与主题相关的 Web 社区。但这个时候链接的立即价值会很小，适合选取基于未来价值的搜索策略，本文采用基于巩固学习的搜索策略。同时为减少网络蜘蛛在爬行过程中的主题漂移，提高主题相关性的判断能力，在每爬完 N 个 Web 社区后（本文用爬行一个固定时间段来表示），系统选取爬虫数据库中爬行到的与主题相关度高前 100 名的页面，与其对应的正向链接信息组成的实例加入链接分类器的训练数据。链接分类器一旦训练完成，就可以对新产生的链接进行相关度分析。自身通过爬虫数据库新进的主题相关度高的页面和页面正向链接信息不断修正，提高主题相关性的判断能力。

2.3 在线增量自适应算法的设计和实现

在线增量自适应算法的本质是：通过网路蜘蛛的爬行，在 Web 社区内尽可能地挖掘和主题相关的页面，而在社区外获取那些具有较高的未来 Q 价值的链接。反过来，在搜索时又根据链接文本的 Q 价值估算出链接的价值，决定选择行动的概率。同时，不断通过爬虫数据库新进的主题相关度高的页面和页面的正向链接信息修正，提高链接主题相关性判断能力。本文利用 Java 技术，算法实现过程如下：

```
ZX-ZL(topic, startUrls){
    Link_1=fetch  link(startUrls);
    While(visited<MAX_PAGES){
                                    //小于爬虫最大访问量
score_r1=sim(topic,  doc);          //计算立即价值
If(socre_r1>r1)
enqueue_1(frontier, extract_links(doc), score_1);
else{
```

```
score_r2=Q(topic, doc);                    //计算未来价值
if(score_2>r2)
continue;
else
enqueue_2(links);
}
}
}
```

2.4 算法过程描述

(1)网络蜘蛛首先从一个“种子集”出发,并选择其中的一个链接访问。

(2)按照式(1)计算链接节点的立即价值。

(3) 判断所得的立即价值是否大于系统给定的阈值r1,如果大于给定阈值,则将该链接加入到候选URL列表里。如果小于给定的阈值r1,就利用式(2)计算此链接的未来价值。

(4) 如果经计算所得未来价值大于系统给定的阈值r2, 系统就并发另一个线程从此节点开始, 返回步骤(2)。

(5)如果所得的未来价值小于给定的阈值r2,将该链接列入被舍弃的URL列表里。结束此线程。

另外每隔 T 的时间后,手动选择与主题相关度高前100的页面加入链接分析器进行训练,对爬虫数据库进行更新[7]。

3 实验与结果分析

3.1 实验背景

本文选取了如表1所示的美国四所大学的计算机网站做了实际的搜索实验,搜索目的是寻找本地服务器中的计算机论文,以PDF和.PS结尾的计算机论文定义为相关文档。采用基于立即价值、未来价值和基于本文所描述的在线增量学习的自适应算法三种不同搜索策略的网络蜘蛛,在线统计Web上与计算机相关的论文数, 并计算各自的查全率和查准率。本文采用FOLDOC在线计算机字典作为主题关键字集合[8]。其中包括13 000个计算机专业词汇, 并进行了一些扩充。从站点的主页出发,对上述四所大学Web服务器进行了实际的搜索测试, 共找到了15 034篇与计算机相关的论文。

表1 四所大学计算机Web服务器名称

大学名称	Web服务器名称
Purdue	www.cs.purdue.edu
Princeton	www.cs.princeton.edu
Brown	www.cs.brown.edu
Marayland	www.cs.marayland.edu

3.2 实验结果和性能分析

图2中,三种不同搜索策略在不同阶段的查全率不同。其原因在于,基于立即价值的搜索策略在相关社区中的搜索率很高,可以很快地找到相关网页,所以其增长率很快。但在找无关网页集合时容易迷失方向,从一个Web社区搜索完毕后进入另一个Web社区的能力较弱,查全率会降低;基于未来价值的搜索策略,在寻找无关页面集合中, 未来价值对预见远期回报很有帮助,它可以很快地找到论文的目录所在, 但早期的回报率不高;基于在线增量自适应算法采用综合的搜索策略,除在搜索初期其回报略低于基于立即价值的网络蜘蛛外,其增长率很快超过两种算法。不论是在社区内的搜索还是过度无关网页来获取远期回报,它都表现出了优异的性能。

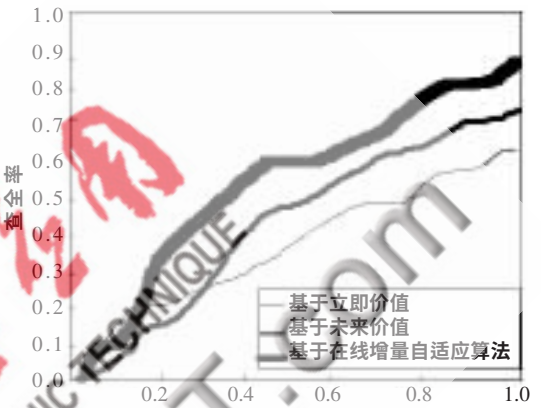

图2 三种蜘蛛的查全率

图3中基于在线增量自适应算法的网络蜘蛛查准率显然高于其他两种。除了最初的阶段外,其余时间的查准率都高于50%。其原因在于每隔一定的时间,爬虫数据库不断自我更新,提高主题相关性的判断能力。在Web社区外, 在一定程度上避免了采集大量的无关文档;在主题相关的Web社区内又提高了其搜索能力,因此其查准率很高。而基于立即价值的网络蜘蛛在跨越Web社区时常常会发生主题偏移, 容易导致局部最优。基于未来价值的网络蜘蛛在跨越Web社区时采集了大量与主题无关的文档,同时在主题相关社区内的搜索能力又比较低,因此查准率不高。

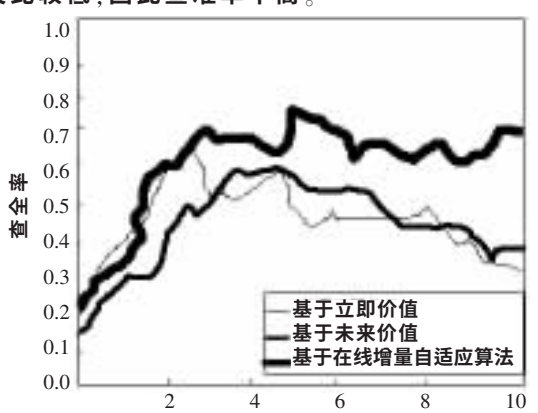

图3 三种蜘蛛的查准率

本文将基于改进的巩固学习方法的行动策略的在线增量自适应算法引入搜索引擎中, 避免了过早陷入

Web 搜索局部最优子空间的陷阱。同时,不断更新爬虫数据库,提高了其对主题相关性的判断能力,从而提高了搜索引擎的查准率。实验表明,该算法的查全率不但大大高于两种传统的单一算法,同时也整体提高了搜索引擎的性能。

参考文献

[1] MURRAY B H, MOORE A. Sizing the internet [M]. A White Paper: Cyveillance, Inc. 2000.

[2] LAWRENCE S, GILES L. Accessibility and distribution of information on the Web[J]. Nature, 1999, 400(8): 107-109.

[3] BREWINGTON B E, CYBENK G. How dynamic is the Web[J]. Computer Networks, 2000, 33(1-6). 257-276.

[4] ESTER M, GROB M, KRIEGEL H. Focused Web crawling: a generic framework for specifying the user interest and for adaptive crawling strategies [Z]. Proceeding of the International Conference on Very Large Database (VLDB'01), 2001.

[5] 陈治平.智能搜索引擎理论与应用研究[D].长沙:湖南大学,2003.

[6] 傅向华,冯博琴,马兆丰,等.可在线增量自学习的聚焦爬行方法[J].西安交通大学学报,2004,38(6):599-602.

[7] HOWE D. Free on-line dictionary of computer[WZ]. http: // www. foldoc. org/. 2010.

[8] CHO J, GARCIA-M H, PAGE L. Efficient crawling through URL ordering[J]. Computer Networks, 1998, 30(1-7): 161-172.

(收稿日期:2010-06-12)

作者简介:

马睿,男,1984 年生,硕士研究生,主要研究方向,人工智能。